ИЗМЕРВАНЕ НА МЕЖДУЕЗИКОВА СЕМАНТИЧНА БЛИЗОСТ ЧРЕЗ ТЪРСЕНЕ В GOOGLE

Светлин Иванов Наков, Софийски университет „Св. Климент Охридски”
email: nakov @ fmi.uni-sofia.bg

Резюме: В настоящата статия е описан алгоритъм за автоматично измерване на семантична близост между двойки думи на различни езици (български и руски). Алгоритъмът извлича локалните контексти на дадените думи чрез серия справки в търсещата машина Google и определя близостта между думите чрез сравнение на контекстите им. За превеждане на контекстите от един език към друг се използва речник. Измерена е корелация от 71% между получените резултати и 30-те двойки думи на Милер и Чарлз, която е по-висока спрямо известните до момента алгоритми.

Ключови думи: Семантична близост, измерване на семантична близост, локален контекст, използване на уеб като корпус, Google.

1. Алгоритъм за извличане на семантична близост чрез търсене в Google

Алгоритъмът изпълнява заявки в търсещата машина Google и анализира върнатите отрязъци от текстове. От тях извлича т. нар. *локален контекст* на всяка анализирана дума (думите в непосредствена близост до нея), тъй като той съдържа думи, които са семантично свързани с нея [Hearst, 1991]. По извлечените локални контексти за всяка дума се построява *честотен вектор*, който съдържа всички думи от съответните локални контексти заедно с честотите им на срещане. Семантичната близост между двойка думи се определя като косинус между честотните им вектори в n-мерното евклидово пространство и представлява число между 0 и 1. Когато разглежданите думи са на различни езици, техните контексти (които също са на различни езици) се сравняват като предварително единият контекст се превежда на другия език чрез речник, както е описано в [Nakov и колектив, 2007a]. Алгоритъмът може да се ползва за измерване на семантична близост не само между думи, но и между фрази.

За извличането на локалния контекст на дадена дума от Интернет използваме заявка за търсене на думата в Google, в която указваме да бъдат върнати 100 резултата на съответния език (в нашия случай български или руски). С 10 такива заявки извличаме до 1000 резултата (Google не позволява да извлечем повече). Всеки резултат съдържа заглавие и отрязък от текст, съдържащи търсената дума или нейна словоформа.

От извлечените резултати първо извличаме всички последователности от думи. Следва премахване на всички функционални думи (предлози, местоимения, съюзи, частици, междуметия и някои наречия), както и думи с по-малко от 3 букви. След това преминаваме през извлечените последователности от думи и търсим дадената дума или нейна словоформа и взимаме 3 думи преди и след нея (числото 3 наричаме размер на контекста). Тези думи считаме за част от локалния уеб контекст. Всички извлечени думи заменяме с тяхната основна словоформа (прилагаме лематизация) като за целта ползваме богати речници на лемите в българския и руския език. Накрая получаваме семантичната близост между двойка думи като пресметнем косинус между честотните им вектори в n-мерното евклидово пространство. Получава се число между 0 и 1, което показва доколко две думи си приличат семантично.

За измерване на семантична близост между думи на различни езици използваме отново контекстите, извлечени от Google, но единият от тях превеждаме на другия език чрез речник от двойки думи, които са превод една на друга. Когато за една дума от единия език има няколко съответни преводни думи от другия език, всяка от тях се взима под

внимание с еднаква тежест. Думите от двата езика, за които няма съответно значение в речника, не се взимат под внимание.

1.1. TF.IDF претегляне

При извличане на информация (information retrieval) често пъти се прилага т. нар. TF.IDF претегляне на честотите на отделните думи, което означава, че по-често срещаните думи участват с по-голяма тежест. Тази техника е описана подробно в [Sparck-Jones, 1972]. Можем да я приложим като преди пресмятане на косинуса между векторите на дадени две думи заместваме честотата на всяка дума от честотния вектор с изчислената за нея TF.IDF стойност.

1.2. Семантична близост, измерена чрез обратен контекст

При извличане на локален контекст за дадена дума от уеб често пъти в него попадат думи, които не са семантично свързани с нея. Премахването на такива думи от локалния уеб контекст следва да доведе до повишаване на точността при оценяване на семантичната близост, защото в контекста ще попадат само думи, които наистина имат семантична връзка с търсената дума [Nakov и колектив, 2007b].

Използването на *обратен контекст* се основава на идеята, че ако две думи са семантично свързани, то първата трябва да се среща често в контекста на втората и същевременно втората трябва да се среща често в контекста на първата. Така честотата на дадена дума A в контекста на друга дума B може да се пресметна два пъти: един път като броя срещания на A в контекста на B и втори път – като броя срещания на B в контекста на A. Накрая може да се вземе по-малката от двете стойности.

При изчисляването на вектора на взаимните срещания чрез обратен контекст е добре да се игнорират думи, които се срещат прекалено малко на брой пъти (примерно по-малко от 10), защото това може да е случайно. С промяна на този параметър (*праг на честотата*) може да се влияе върху точността на резултатите.

1.3. Семантична близост чрез обогатяване на контекста

Обогатяване на контекста означава да добавим към контекста на дадена дума контекстите на всички често срещани в него думи [Hagiwara и колектив, 2007]. По този начин контекстът на думата се разширява с още думи, които оригинално не присъстват в него, но са свързани смислово с тази дума. Очакванията са това да подобри точността на алгоритъма за измерване на семантичната близост между двойка думи.

При обогатяване на контекста, е добре да се игнорират думи, които се срещат в него прекалено малко на брой пъти (примерно по-малко от 10), защото това може да е случайно. С промяна на този параметър (*праг на честотата*) може да се влияе върху точността на резултатите, като се зададе разумна граница на минималния брой срещания, при който се извършва обогатяване на контекста.

2. Експерименти и резултати

Експериментите, които направихме, имат за цел да оценят предложените алгоритми за автоматично извличане на междуезикова семантична близост от уеб чрез сравнение на получените от тях резултати с оценки, дадени от човек.

2.1. Тестови данни

Като тестови данни използваме списъка от 30 двойки думи, предложени от Милер и Чарлз [Miller & Charles, 1991]. Те представляват внимателно подбрани двойки съществителни имена, за всяка от които е направена оценка на семантичната близост от 51 души в скала от 0 до 4, след което оценката е усреднена. Посочените от Милер и

Чарлз 30 двойки думи преведохме съответно на български и руски език. При превода не навсякъде успяхме да намерим точно съответствие между английски, български и руски език и на много места се изгубиха нюансите на оригиналните думи, което би могло да направи неточна дадената от Милер и Чарлз човешка оценка, но ние приемаме този риск и знаем, че не можем да очакваме 100% точност на резултатите.

2.2. Използвани ресурси

За целите на експериментите и при реализирането на алгоритъма за извличане на семантична близост от уеб бяха използвани следните ресурси:

- **Граматичен речник на българския и руския език** [Paskaleva, 2007]. В българския си вариант речникът съдържа 963 339 словоформи и 73 113 леми. В руския си вариант речникът съдържа 1 390 613 словоформи и 66 101 леми.
- **Списък с функционалните думи в българския и руския език** (598 български и 507 руски думи: предлози, местоимения, съюзи, частици, междуметия и наречия).
- **Кратък българо-руски речник**, съдържа 4 562 двойки думи, които са превод една на друга. Съставен от онлайн българо-руски речник [BgRu.net, 2007].
- **Подробен българо-руски речник**, съдържа 59 582 двойки думи и фрази, които са превод една на друга. Съставен от два големи българо-руски и руско-български речника [Чукалов, 1986] и [Бернщайн, 1986].

2.3. Описание на експериментите

Върху адаптираните от Милер и Чарлз 30 двойки думи и фрази са проведени серия експерименти за оценяване на семантичната им близост чрез изпълнение на описаните алгоритми при различни техни параметри (с различна големина на речниците, с и без прилагане на TF.IDF, с и без използване на обратен контекст, с и без обогатяване на контекста и при различни стойности на минималната честота на срещане на думите):

- **RAND** – случайна близост, зададена за всички двойки думи.
- **SIM** – основният алгоритъм за извличане на семантична близост от уеб.
- **SIM-BIG** – основният алгоритъм SIM с подробния българо-руски речник.
- **SIM+TFIDF** – модификация на SIM алгоритъма с използване на TF.IDF.
- **SIM-BIG+TFIDF** – модификация на SIM алгоритъма с използване на TF.IDF претегляне и използване на подробния българо-руски речник.
- **REV-0**, **REV-10**, **REV-20**, **REV-30**, **REV-40**, **REV-50** – модификация на SIM алгоритъма с използване на обратен контекст с прагове 0, 10, 20, 30, 40 и 50.
- **REV-BIG-0**, **REV-BIG-10**, **REV-BIG-20**, **REV-BIG-30**, **REV-BIG-40**, **REV-BIG-50** – модификация на REV алгоритъма с подробния българо-руски речник.
- **IND-10**, **IND-20**, **IND-30**, **IND-40**, **IND-50** – модификация на SIM алгоритъма с използване на обогатен контекст с прагове 10, 20, 30, 40 и 50.
- **IND-BIG-10**, **IND-BIG-20**, **IND-BIG-30**, **IND-BIG-40**, **IND-BIG-50** – модификация на IND алгоритъма с използване на подробния българо-руски речник.

2.4. Резултати

Получените резултати от нашите алгоритми за автоматичното измерване на семантична близост са сравнени с човешката оценка чрез изчисление на *коефициента на корелация на Пирсън* (стандартна статистическа мярка за измерване на линейна взаимовръзка между две променливи величини). Получени са следните резултати:

Алгоритъм	Праг 0	Праг 10	Праг 20	Праг 30	Праг 40	Праг 50
RAND	**0,0000**	-	-	-	-	-
SIM	**0,7043**	-	-	-	-	-
SIM+TFIDF	**0,7010**	-	-	-	-	-
SIM-BIG	**0,6210**	-	-	-	-	-
SIM-BIG+TFIDF	**0,6191**	-	-	-	-	-
REV	**0,5933**	0,5732	0,5623	0,5625	0,5623	0,5492
REV-BIG	0,5961	**0,5964**	0,5956	0,5957	0,5953	0,5920
IND	-	0,5078	0,6027	**0,6850**	0,6485	0,6445
IND-BIG	-	0,5046	0,6057	**0,7149**	0,6296	0,6412

2.5. Анализ на резултатите

От таблицата е видно, че семантичната близост, оценена автоматично с предложените алгоритми, има корелация със съответните човешки оценки от 50% до 71%. Тази корелация е много по-висока от 0%, която се получава при случайната оценка RAND.

Макар и резултатите от основния алгоритъм SIM да са доста добри, виждаме, че всички опити за неговото подобрение не са много успешни. От резултатите можем да направим следните заключения:

- Използването на TF.IDF претегляне влияе негативно.
- Използването на обратен контекст не работи добре и REV алгоритъма работи по-лошо от основния SIM алгоритъм при всякакви прагове на честотата.
- Обогатяването на контекста (IND алгоритъма) работи малко по-добре от основния SIM алгоритъм само при внимателно подбран праг на честотата.
- Използването на подробния вместо краткия речник помага само в някои случаи.

Основните причини за неточност на резултатите са няколко:

- Загуба на нюанси при превода на 30-те думи на Милер и Чарлз.
- Използването на уеб като корпус ограничава извличането на локален контекст измежду само 1000 статии, а те не са представителна извадка на всички статии.
- Използването на думи, а не фрази, при извличане на контекстите и след това при превода внася много шум. Това е основен недостатък на описаните алгоритми.
- Непълнота на преводните речници. Думите от двата езика, за които няма съответно значение в речника, не се взимат под внимание (игнорират се).

3. Други разработки по темата и сравнение с тях

Повечето известни методи за автоматично оценяване на семантична близост се базират на лингвистичната хипотеза за разпределението (distributional hypothesis) [Harris, 1954], която твърди, че семантично близките думи се срещат в близки контексти. На нея е основан и нашият подход.

[Weeds, 2003] сравнява 6 алгоритъма за извличане на семантична близост, базирани на хипотезата за разпределението (и неизползващи допълнителни ресурси), и установява, че най-добрият от тях постига коефициент на корелация на Пирсън от 62%. Нашият най-добър резултат от 71% е по-добър от описаните от Weeds алгоритми, като при това се отнася за по-сложната задача за измерване междуезикова семантична близост.

[Budanitsky & Hirst, 2006] сравняват 5 различни алгоритъма за оценяване на семантична близост, базирани на WordNet. Най-добрият от тях постига корелация на Пирсън от 85% за 30-те двойки думи на Милер и Чарлз, което е по-добър резултат в сравнение с нашите алгоритми, но използва WordNet (който няма български и руски вариант).

4. Библиография

[Hearst, 1991] Hearst M. "Noun Homograph Disambiguation Using Local Context in Large Text Corpora". *7th Annual Conference of the University of Waterloo*, Oxford, 1991.

[Nakov и колектив, 2007a] Nakov P., Nakov S., Paskaleva E. “Improved Word Alignments Using the Web as a Corpus”. *Proc. of RANLP'2007*, Bulgaria, 2007.

[Nakov и колектив, 2007b] Nakov S., Nakov P., Paskaleva E. “Cognate or False Friend? Ask the Web!”. *Workshop on Acquisition and Management of Multilingual Lexicons,* Bulgaria, 2007.

[Sparck-Jones, 1972] Sparck-Jones K. “A Statistical Interpretation of Term Specificity and its Application in Retrieval”. *Journal of Documentation*, Volume 28, 1972.

[Miller & Charles, 1991] Miller, G., Charles W., Contextual Correlates of Semantic Similarity, *Language and Cognitive Processes*, 1991, 6(1):1–28

[Hagiwara и колектив, 2007] Hagiwara M., Ogawa Y., Toyama K. (2007). "Effectiveness of Indirect Dependency for Automatic Synonym Acquisition". *CoSMo 2007 Workshop,* Denmark, 2007.

[Paskaleva, 2002] Paskaleva E. “Processing Bulgarian and Russian Resources in Unified Format”. *8th International Scientific Symposium MAPRIAL*, Bulgaria, 2002.

[BgRu.net, 2007] Online Bulgarian-Russian dictionary – http://www.bgru.net/intr/dictionary/.

[Чукалов, 1986] Чукалов С. К. "Руско-български речник", Изд. "Русски език", Москва, 1986.

[Бернщайн, 1986] Бернщайн, С. Б. "Българо-руски речник". Изд. "Русски език", Москва, 1986.

[Harris, 1954] Harris, Z. "Distributional structure”. *Word*, volume 10, 1954.

[Weeds, 2003] Weeds J. “Measures and Applications of Lexical Distributional Similarity”, *Ph.D. Thesis*, University of Sussex, 2003

[Budanitsky & Hirst, 2006] Budanitsky A., Hirst G. ”Evaluating WordNet-based Measures of Lexical Semantic Relatedness”. *Computational Linguistics*, Volume 32, 2006.